Cambridge Handbook of Experimental Political Science

Edited by

JAMES N. DRUCKMAN
Northwestern University

DONALD P. GREEN
Yale University

JAMES H. KUKLINSKI
University of Illinois at Urbana–Champaign

ARTHUR LUPIA
University of Michigan
INTROD

1 Exper
James

PART I:

2 Exper
James

3 Inte
Rose N

4 Stude
James

5 Eco
and L
Eric S

PART II:
IN POLI'

6 Labor
Shanti
CHAPTER 4

Students as Experimental Participants

A Defense of the "Narrow Data Base"

James N. Druckman and Cindy D. Kam

An experiment entails randomly assigning participants to various conditions or manipulations. Given common consent requirements, this means experimenters need to recruit participants who, in essence, agree to be manipulated. The ensuing practical and ethical challenges of subject recruitment have led many researchers to rely on convenience samples of college students. For political scientists who put particular emphasis on generalizability, the use of student participants often constitutes a critical, and according to some reviewers, fatal problem for experimental studies.

In this chapter, we investigate the extent to which using students as experimental participants creates problems for causal inference. First, we discuss the impact of student subjects on a study's internal and external validity. In contrast to common claims, we argue that student subjects do not intrinsically pose a problem for a study's external validity. Second, we use simulations to identify situations when student subjects are likely to constrain experimental inferences. We show that such situations are relatively limited; any convenience sample poses a problem only when the size of an experimental treatment effect depends on a characteristic on which the convenience sample has virtually no variance. Third, we briefly survey empirical evidence that provides guidance on when researchers should be particularly attuned to taking steps to ensure appropriate generalizability from student subjects. We conclude with a discussion of the practical implications of our findings. In short, we argue that student subjects are not an inherent problem to experimental research; moreover, the burden of proof — articulating and demonstrating that student subjects pose an inferential problem — should lie with critics rather than experimenters.

1. The "Problem" of Using Student Subjects

Although internal validity may be the sine qua non of experiments, most researchers use experiments to make generalizable causal
inferences (Shadish, Cook, and Campbell 2002, 18–20). For example, suppose one implements a laboratory study with students and finds a causal connection between the experimental treatment (e.g., a media story about welfare) and an outcome of interest (e.g., support for welfare). An obvious question is whether the relationship found in the study exists within a heterogeneous population, in various contexts (e.g., a large media marketplace), and over time. This is an issue of external validity, which refers to the extent to which the “causal relationship holds over variations in persons, settings, treatments [and timing], and outcomes” (Shadish et al. 2002, 83). McDermott (2002) explains, “External validity . . . tend[s] to preoccupy critics of experiments. This near obsession . . . tend[s] to be used to dismiss experiments” (334).

A point of particular concern involves generalization from the sample of experimental participants – especially when, as is often the case, the sample consists of students – to a larger population of interest. Indeed, this was the focus of Sears’ (1986) widely cited article, “College Sophomores in the Laboratory: Influences of a Narrow Data Base on Social Psychology’s View of Human Nature.” Many political scientists simply assume that “a student sample lacks external generalizability” (Kam, Wilking, and Zechmeister 2007, 421). Gerber and Green (2008) explain, “If one seeks to understand how the general public responds to social cues or political communication, the external validity of lab studies of undergraduates has inspired skepticism” (Sears 1986; Benz and Metier 2008, 358). In short, social scientists in general and political scientists in particular view student subjects as a major hindrance to drawing inferences from experimental studies.

Assessing the downside of using student subjects has particular relevance. First, many political science experiments use student subjects; for example, Kam et al. report that from 1990 through 2006, one fourth of experimental articles in general political science journals relied on student subjects, whereas more than seventy percent did so in more specialized journals (Kam et al. 2007, 419–20; see also Druckman et al. 2006). Are the results from these studies of questionable validity? Second, there are practical issues. A common rationale for moving away from laboratory studies – in which student subjects are relatively common – to survey and/or field experiments is that these latter venues facilitate using nonstudent participants. When evaluating the pros and cons of laboratory versus survey or field experiments, should substantial weight be given to whether participants are students? Similarly, those implementing lab experiments have increasingly put forth efforts (and paid costs) to recruit nonstudent subjects (e.g., Lau and Redlawsk 2006, 65–66; Kam 2007). Are these costs worthwhile? To address these questions, we next turn to a broader discussion of what external validity demands.

Dimensions of External Validity

To assess the external validity or generalizability of a causal inference, one must consider from what we are generalizing and to what we hope to generalize. When it comes to “from what,” a critical, albeit often neglected, point is that external validity is best understood as being assessed over a range of studies on a single topic (McDermott 2002, 335). Assessment of any single study, regardless of the nature of its participants, must be done in light of the larger research agenda to which it hopes to contribute.\footnote{This is consistent with a Popperian approach to evaluation that suggests causal hypotheses are never confirmed and evidence accumulates via multiple tests, even if these tests have limitations. Campbell (1969) offers a fairly extreme stance on this when he states, “If we achieved one, there would be no need to apologize for a successful psychology of college sophomores, or even of Northwestern University coeds, or of Watertown white rats” (961).}

Moreover, when it comes to generalization from a series of studies, the goal is to generalize across multiple dimensions. External
validity refers to generalization not only of individuals, but also across settings/contexts, times, and operationalizations. There is little doubt that institutional and social contexts play a critical role in determining political behavior and that, consequently, they can moderate causal relationships. One powerful recent example comes from the political communication literature; a number of experiments, using both student and non-student subjects, show that when exposed to political communications (e.g., in a laboratory), individuals’ opinions often reflect the content of those communications (e.g., Kinder 1998; Chong and Druckman 2007b). The bulk of this work, however, ignores the contextual reality that people outside the controlled study setting have choices (i.e., they are not captive). Arceneaux and Johnson (2008) show that as soon as participants in communication experiments can choose whether to receive a communication (i.e., the captive audience constraint is removed), results about the effects of communications drastically change (i.e., the effects become less dramatic). In this case, ignoring the contextual reality of choice appears to have constituted a much greater threat to external validity than the nature of the subjects.3

Timing also matters. Results from experiments implemented at one time may not hold at other times, given the nature of world events. Gaines, Kuklinski, and Quirk (2007) further argue that survey experiments in particular may misestimate effects due to a failure to consider what happened prior to the study (also see Gaines and Kuklinski’s chapter in this volume). Building on this insight, Druckman (2009) asked survey respondents for their opinions about a publicly owned gambling casino, which was a topic of “real-world” ongoing political debate. Prior to expressing their opinions, respondents randomly received no information (i.e., control group) or information that emphasized either economic benefits or social costs (e.g., addiction to gambling). Druckman shows that the opinions of attentive respondents (i.e., respondents who regularly read newspaper coverage of the campaign) in the economic information condition did not significantly differ from attentive individuals in the control group.4 The noneffect likely stemmed from the economic information—which was available outside the experiment in ongoing political discussion—having already influenced all respondents. Another exposure to this information in the experiment did not add to the prior, pre-treatment effect. In other words, the ostensible noneffect lacked external validity, not because of the sample but because it failed to account for the timing of the treatment and what had occurred prior to that time (also see Slothuus 2009).5

A final dimension of external validity involves how concepts are employed. Finding support for a proposition means looking for different ways of administering and operationalizing the treatment (e.g., delivering political information via television ads, newspaper stories, interpersonal communications, survey question text) and operationalizing the dependent variables (e.g., behavioral, attitudinal, physiological, implicit responses).

In short, external validity does not simply refer to whether a specific study, if rerun on a different sample, would provide the same results. It refers more generally to whether “conceptually equivalent” (Anderson and Bushman 1997, 21) relationships can be detected across people, places, times, and operationalizations. This introduces the other end of the generalizability relationship—that is, “equivalent” to what? For many,

---
3 A related example comes from Barabas and Jerit’s (2010) study, which compares the impact of communications in a survey experiment against analogous dynamics that occurred in actual news coverage. They find that the survey experiment vastly overstated the effect, particularly among certain subgroups. Sniderman and Theriault (2004) and Chong and Druckman (2009a) also reveal the importance of context; both studies show that prior work that limits competition between communications (i.e., by only providing participants with a single message rather than a mix that is typically found in political contexts) likely misestimates the impact of communications on public opinion.

4 For reasons explained in his paper, Druckman (2009) also focuses on individuals more likely to have formed prior opinions about the casino.

5 Another relevant timing issue concerns the duration of any experimental treatment effect (see, e.g., Gaines et al. 2007; Gerber et al. 2007).
the "to what" refers to behavior as observed outside the study, but this is not always the case. Experiments have different purposes; Roth (1995) identifies three nonexclusive roles that experiments can play: search for facts, speaking to theorists, or whispering in the ears of princes, (22) which facilitates "the dialogue between experimenters and policymakers" (see also Guara 2005, 141–60). These types likely differ in the target of generalization. Of particular relevance is that theory-oriented experiments are typically not meant to "match" behaviors observed outside the study per se, but rather the key is to generalize to the precise parameters put forth in the given theory. Plott (1991) explains, "The experiment should be judged by the lessons it teaches about the theory and not by its similarity with what nature might have happened to have created" (p.6). This echoes Mook's (1983) argument that much experimental work is aimed at developing and/or testing a theory, not at establishing generalizability. Experiments that are designed to demonstrate "what can happen" (e.g., Milgram 1963; Zimbardo 1973) can still be useful, even if they do not mimic everyday life.6 In many instances, the nature of the subjects in the experiments are of minimal relevance, particularly given experimental efforts to ensure that their preferences and/or motivations match those in the theory (e.g., see Dickson’s chapter in this volume).

Assessment of how student subjects influence external validity depends on three considerations: 1) the research agenda on which the study builds (e.g., has prior work already established a relationship with student subjects, meaning incorporating other populations may be more pressing?); 2) the relative generalizability of the subjects, compared to the setting, timing, and operationalizations (e.g., a study using students may have more leeway to control these other dimensions); and 3) the goal of the study (e.g., to build a theory or to generalize one).

Evaluating External Validity

The next question is how to evaluate external validity. Although this is best done over a series of studies, we acknowledge the need to assess the strengths of a particular study with respect to external validity. Individual studies can be evaluated in at least two ways (Aronson and Carlsmith 1968; Aronson, Brewer, and Carlsmith 1985; Aronson, Wilson, and Brewer 1990). First, experimental realism refers to whether "an experiment is realistic, if the situation is involving to the subjects, if they are forced to take it seriously, [and] if it has impact on them" (Aronson, Brewer, and Carlsmith 1985, 485). Second, mundane realism concerns "the extent to which events occurring in the research setting are likely to occur in the normal course of the subjects' lives, that is, in the 'real world'" (Aronson et al. 1985, 485).

Much debate about samples focuses on mundane realism. When student subjects do not match the population to which a causal inference is intended, many conclude that the study has low external validity. Emphasis on mundane realism, however, is misplaced (e.g., McDermott 2002; Morton and Williams 2008, 345); of much greater importance is experimental realism. Failure of participants to take the study and treatments "seriously" compromises internal validity, which in turn renders external validity of the causal relationship meaningless (e.g., Dickhaut, Livingstone, and Watson 1972, 477; Liyanarachchi 2007, 56).8 In contrast, at worst, low levels

---

6 Aronson, Wilson, and Brewer (1998) explain that it "is often assumed (perhaps mindlessly) that all studies should be as high as possible in external validity, in the sense that we should be able to generalize the results as much as possible across populations and settings and time. Sometimes, however, the goal of the research is different" (131).

7 A third evaluative criterion is psychological realism, which refers to "the extent to which the psychological processes that occur in an experiment are the same as psychological processes that occur in everyday life" (Aronson et al. 1998, 132). The relevance of psychological realism is debatable and depends on one's philosophy of science (c.f. Friedman 1953; Simon 1963, 1979, 473–76; also see MacDonald 2003).

8 By "seriously," we mean analogous to how individuals treat the same stimuli in the settings to which one hopes to generalize (and not necessarily "serious" in a technical sense). We do not further discuss steps that can be taken to ensure experimental realism because this moves into the realm of other design issues.
of mundane realism simply constrain the breadth of any generalization but do not make the study useless.

Moreover, scholars have yet to specify clear criteria for assessing mundane realism, and, as Liyanarachchi (2007) explains, “Any superficial appearance of reality (e.g., a high level of mundane realism) is of little comfort, because the issue is whether the experiment ‘captures the intended essence of the theoretical variables’” (Kruglanski 1975, 106) (57). That said, beyond superficiality, we recognize that student subjects — although having no ostensibly relevant connection with experimental realism — may limit mundane realism that constrains generalizations of a particular study. This occurs when characteristics of the subjects affect the nature of the causal relationships being generalized. When this happens, and with what consequences, are questions to which we now turn.

### 2. Statistical Framework

In this section, we examine the use of student samples from a statistical point of view. This allows us to specify the conditions under which student samples might constraining causal generalization (in the case of a single experiment). Our focus, as in most political science analyses of experimental data, is on the magnitude of the effect of some experimental treatment, $T$, on an attitudinal or behavioral dependent measure, $y$. Suppose, strictly for presentational purposes, we are interested in the effect of a persuasive communication ($T$) on a subject’s poststimulus policy opinion ($y$). $T$ takes on a value of 0 for subjects randomly assigned to the control group and takes on a value of 1 for subjects randomly assigned (e.g., subject payments, incentives; see Dickson’s chapter in this volume).

9 Berkowitz and Donnerstein (1981) explain, “The meaning the subjects assign to the situation they are in and the behavior they are carrying out [i.e., experimental realism] plays a greater part in determining generalisability of an experiment’s outcome than does the sample’s demographic representatives or the setting’s surface realism” (240).

10 This claim is in need of empirical evaluation because it may be that students are more compliant, and this may have an impact on realism.

11 Suppose the true data-generating process features a homogeneous treatment effect:

$$y_i = \beta_0 + \beta_T T_i + \epsilon_i.$$  

(1)

Assuming that all assumptions of the classical linear regression model are met, the ordinary least squares estimate for $\beta_T$ is unbiased, consistent, and efficient. The results derived from estimation on a given sample would be fully generalizable to those that would result from estimation on any other sample.

Specific samples will differ in their distributions of individual covariates. Continuing with our running example, samples may differ in the distribution of attitude crystallization (i.e., an attitude is increasingly crystallized when it is stronger and more stable). Student samples may yield a disproportionately large group of subjects who are low in crystallization. A random sample from the general population might generate a group that is normally distributed and centered at the

12 For ease of exposition, our example only has one treatment group. The lessons easily extend to multiple treatment groups.

13 We could have specified a data-generating process that also includes a direct relationship between $y$ and some individual-level factors such as partisanship or sex (consider a vector of such variables, X). Under random assignment, the expected covariance between the treatment and X is zero. Hence, if we were to estimate the model without X, then omitted variable bias would technically not be an issue. If the data-generating process does include X, and even though we might not have an omitted variable bias problem, including X in the model may still be advisable. Inclusion of relevant covariates (i.e., covariates that, in the data-generating process, actually have a nonzero effect on $y$) will reduce $\epsilon$ (the difference between the observed and predicted $y$), which in turn will reduce $\sigma^2$, resulting in more precise estimated standard errors for our coefficients (see Franklin 1997). Moreover, it is only in expectation that Cov(X,T) = 0. In any given sample, Cov(X,T) may not equal zero. Inclusion of covariates can mitigate against incidental variation in cell composition. In advising inclusion of control variables, Ansolabehere and Iyengar (1995) note, “Randomization does not always work. Random assignment of treatments provides a general safeguard against bias but it is not foolproof. By chance, too many people of a particular type may end up in one of the treatment groups, which might skew the results” (172; see also Bowers’ chapter in this volume).

14 This example is inspired by Sears’ (1986) discussion of “Uncrystallized Attitudes.”
middle of the range. A sample from politically active individuals (e.g., conventioneers) might result in a group that is disproportionately high in crystallization.14

Consider the following samples with varying distributions on attitude crystallization. In all cases, n = 200, and treatment is randomly assigned to half the cases. Attitude crystallization ranges from 0 (low) to 1 (high). Consider a “student sample,” where ninety percent of the sample is at a value of “0” and ten percent of the sample is at a value of “1.” Consider a “random sample,” where the sample is normally distributed and centered on 0.5, with standard deviation of 0.165. Finally, consider a “conventioneers sample,” where ten percent of the sample is at a value of “0” and ninety percent of the sample is at a value of “1.”15

Suppose the true treatment effect (βτ) takes a value of “4.” We set up a Monte Carlo experiment that estimated Equation (1) 1,000 times, each time drawing a new e term. We repeated this process for each of the three types of samples (student, random, and conventioneers). The sampling distributions for βτ appear in Figure 4.1.

The results demonstrate that when the true data-generating process produces a single treatment effect, estimates from any sample will produce an unbiased estimate of the true underlying treatment effect. Perhaps this point seems obvious, but we believe it has escaped notice from many who criticize experiments that rely on student samples. We repeat: if the underlying data-generating process is characterized by a heterogeneous treatment effect (i.e., the treatment effect is the same across the entire population), then any convenience sample should produce an unbiased estimate of that single treatment effect, and, thus, the results from any convenience sample should generalize easily to any other group.

Suppose, however, the “true” underlying data-generating process contains a heterogeneous treatment effect: that is, the effect of the treatment is moderated16 by individual-level characteristics. The size of the treatment effect might depend on some characteristic, such as gender, race, age, education, sophistication, etc. Another way to say this is that there may be an “interaction of causal relationship with units” (Shadish et al. 2002, 87).

As one method of overcoming this issue, a researcher can randomly sample experimental subjects. By doing so, the researcher can be assured that

Although random sampling has advantages for external validity, Shadish et al. note that “it is so rarely feasible in experiments” (91). The way to move to random sampling might be to use survey experiments, where respondents are (more or less) a random sample of some population of interest. We say a bit more about this possibility later in the chapter. For now, let us assume that a given researcher has a specific set of reasons for not using a random sample (cost, instrumentation, desire for laboratory control, etc.), and let’s examine the challenges a researcher

14 And, of course, crystallization might vary across different types of issues. On some issues (e.g., financial aid policies), students might have highly crystallized views, whereas conventioneers might have less crystallized views.

15 Now, if our goal was to use our three samples to make descriptive inferences about the general population’s mean level of attitude crystallization, then both the student sample and the conventioneers sample would be inappropriate. The goal of an experimental design is expressly not to undertake this task. Instead, the goals of an experimental design are to estimate the causal effect of some treatment and then to generalize it.

16 See Baron and Kenny (1986) for the distinction between moderation and mediation. Psychologists refer to the case where Z affects the effect of X as moderation (i.e., an interaction effect). Psychologists refer to mediation when some variable X influences the level of some variable Z, whereby X influences the level on the level of Z. For an extended treatment of interaction effects in regression analysis, see Kam and Franzese (2007). For a discussion of mediation, see Bullock and Ha’s chapter in this volume.
mate of that single result from any 
realize easily to any 

rue" underlying a heteroge-
is, the effect of 6 by individual-
of the treatment e characteristic,
ucation, sophis-
say this is that of causal relat-
ing this issue, a 
ple experiment-

p observed in 
(t) the aver-
uld have been 
ample population 
ship that 
other persons 
et al. 2002,

as advantages 
t al. note that 
timents" (91).
pling might 
repondom 
sit. We say a 
 later in the 
 at a given 
 f reasons for 
t, instrument-
, etc.), 
 researcher 

the distinction 
Psychologists 
not of X as mod-
ologists refer 
mes the 
nded treatment 
alysis, see Kam 
 n of mediation, 

using a convenience sample might face in this framework.

We revise our data-generating process to reflect a heterogeneous treatment effect by taking Equation (1) and modeling how some individual-level characteristic, Z (e.g., attitude crystallization), influences the magnitude of the treatment effect:

$$\beta_1 = \gamma_{10} + \gamma_{11}Z_i.$$  \hfill (2)

We also theorize that Z might influence the intercept:

$$\beta_0 = \gamma_{00} + \gamma_{01}Z_i.$$  \hfill (3)

If our sample includes sufficient variance on 
this moderator, and we have ex ante theo-
ized that the treatment effect depends on 
this moderating variable, Z, then we can (and should) estimate the interaction. If, however, 
the sample does not contain sufficient var-
ance, not only can we not identify the moder-
ating effect, but we may also misestimate 
the on-average effect depending on what specific range of Z is present in our sample.

The question of generalizing treatment effects reduces to asking whether there is a single treatment effect or a set of treatment effects, the size of which depends on some (set of) covariate(s). Note that this is a theoretically oriented question of generalization. It is not 
just whether "student samples are generaliz-
able," but rather what particular characteristics of student samples might lead us to won-
der whether the causal relationship detected in a student sample experiment would be system-
tically different from the causal relation-
ship in the general population.
Revisiting our running example, suppose we believe that a subject’s level of attitude crystallization ($Z$) influences the effect of a persuasive communication ($T$) on a subject’s poststimulus policy opinion ($y$). The more crystallized someone’s attitude is, the smaller the treatment effect should be, whereas the less crystallized someone’s attitude is, the greater the treatment effect should be. Using this running example, based on Equation (3), assume that the true relationship has the following (arbitrarily selected) values:

\[
\begin{align*}
\gamma_{00} &= 0 \\
\gamma_{01} &= 0 \\
\gamma_{10} &= 5 \\
\gamma_{11} &= -5
\end{align*}
\]

Let $Z$, attitude crystallization, range from 0 (least crystallized) to 1 (most crystallized). $\gamma_{10}$ tells us the effect of the treatment when $Z = 0$, that is, the treatment effect among the least crystallized subjects. $\gamma_{11}$ tells us how crystallization moderates the effect of the treatment.

First, consider what happens when we estimate Equation (1), the simple (but theoretically incorrect, given that it fails to model the moderating effect) model that looks for the “average” treatment effect. We estimated this model 1,000 times, each time drawing a new $\varepsilon$ term. We repeated this process for each of the three samples. The results appear in Figure 4.2.

When we estimate a “simple” model looking for an average treatment effect, our estimates for $\beta_1$, diverge from sample to sample. In cases where we have a student sample, and where low levels of crystallization increase the treatment effect, we systematically overestimate the treatment effect relative to what we would get in estimating the same model on a random sample with moderate levels of crystallization. In the case of a conventioneers sample, where high levels of crystallization depress the treatment effect, we systematically underestimate the treatment effect, relative to the estimates obtained from the general population.

We have obtained three different results across the samples because we estimated a model based on Equation (1). Equation (1) should only be estimated when the data-generating process produces a single treatment effect: the value of $\beta_1$. However, we have “mistakenly” estimated Equation (1) when the true data-generating process produces a series of treatment effects (governed by the function $\beta_1 = 5 - 5Z$). The sampling distributions in Figure 4.2 provide the “average” treatment effect, which depends directly on the mean value of $Z$ within a given sample: $5 - 5E(Z)$.

Are the results from one sample more trustworthy than the results from another sample? As Shadish et al. (2002) note, conducting an experiment on a random sample will produce an “average” treatment effect; hence, to some degree, the results from the random sample might be more desirable than the results from the other two convenience samples. However, all three sets of results reflect a fundamental disjuncture between the model that is estimated and the true data-generating process. If we have a theoretical reason to believe that the data-generating process is more complex (i.e., the treatment depends on an individual-level moderator), then we should embed this theoretical model into our statistical model.

To do so, we returned to Equation (3) and estimated the model 1,000 times, each time drawing a new $\varepsilon$ term. We repeated this process three times, for each of the three samples. The results appear in Figure 4.3.

First, notice that the sampling distributions for $\beta_T$ are all centered on the same value, 5, and the sampling distributions for $\beta_{TZ}$ are also all centered on the same value, $-5$. In other words, Equation (3) produces unbiased point estimates for $\beta_T$ and $\beta_{TZ}$, regardless of which sample is used. We uncover unbiased point estimates even where only ten percent of the sample provides key variation on $Z$ (student sample and conventioneers sample).

Second, notice the spread of the sampling distributions. We have the most certainty about $\beta_T$ in the student sample and substantially less certainty in the random sample and the conventioneers sample. The greater degree of certainty in the student sample results from the greater mass of the sample.
n (1). Equation (1) is modified when the data suggests a single treatment effect, $\beta$, However, we noted Equation (1) rating process produces a single treatment effect (governed by $Z_0$). The sampling process provides the "average" depends directly on the sample size.

One sample more results from another (2002) note, common a random sample 3" treatment effect; the results from the more desirable than the other two convenience sets of results are more uncertain between the data and the true data. We have a theoretical data-generating (i.e., the treatment at-level moderator), theoretical model to Equation (3) and 300 times, each time we repeated this process of the three samples.

Figure 4.2. Sampling Distribution of $b_T$, Heterogeneous Treatment Effects

Note: 1,000 iterations, estimated using Equation (1).

![Figure 4.2](image)

Figure 4.3. Sampling Distributions of $b_T$ and $b_{Z_2}$, Heterogeneous Treatment Effects

Note: 1,000 iterations, estimated using Equation (3).
that is located at o in the student sample (because the point estimate for \( b_T \), the uninteracted term in Equation (3), represents the effect of \( T \) when \( Z \) takes on the value of o).

For the sampling distribution of \( b_{TZ} \), we have higher degrees of certainty (smaller standard errors) in the student sample and the convention-seekers sample. This is an interesting result. By using samples that have higher variation on \( Z \), we yield more precise point estimates of the heterogeneous treatment effect. Moreover, we are still able to uncover the interactive treatment effect because these samples still contain some variation across values of \( Z \).

How much variation in \( Z \) is sufficient? As long as \( Z \) varies to any degree in the sample, the estimates for \( b_T \) and \( b_{TZ} \) will be unbiased. Being “right on average” may be of little comfort if the degree of uncertainty around the point estimate is large. If \( Z \) does not vary very much in a given sample, then the estimated standard error for \( b_{TZ} \) will be large. But concerns about uncertainty are “run of the mill” when estimating a model on any dataset: more precise estimates arise from analyzing datasets that maximize variation in our independent variables.

Our discussion thus suggests that experimentalists (and their critics) need to consider the underlying data-generating process: that is, theory is important. If a single treatment effect is theorized, then testing for a single treatment effect is appropriate. If a heterogeneous treatment effect is theorized, then researchers should explicitly theorize how the treatment effect should vary along a specific (set of) covariate(s), and researchers can thereby estimate such relationships as long as there is sufficient variation in the specific (set of) covariate(s) in the sample. We hope to push those who launch vague criticisms regarding the generalizability of student samples to offer more constructive, more theoretically oriented critiques that reflect the possibility that student samples may be problematic if the magnitude and direction of the treatment effect depend on a particular (set of) covariate(s) that are peculiarly distributed within a student sample.

In sum, we have identified three distinct situations. First, in the homogeneous case — where the data-generating process produces a single treatment effect — we showed that the estimated treatment effect derived from a student sample is an unbiased estimate of the true treatment effect. Second, when there is a heterogeneous case (where the treatment effect is moderated by some covariate \( Z \)) and the researcher fails to recognize the contingent effect, a student sample (indeed, any convenience sample) may misestimate the average treatment effect if the sample is non-representative on the particular covariate \( Z \). However, in this case, even a representative sample would misspecify the treatment effect due to a failure to model the interaction. Third, when the researcher appropriately models the heterogeneity with an interaction, then the student sample, even if it is nonrepresentative on the covariate \( Z \), will misestimate the effect only if there is virtually no variance (i.e., literally almost none) on the moderating dynamic. Moreover, a researcher can empirically assess the degree of variance on the moderator within a given sample and/or use simulations to evaluate whether limited variance poses a problem for uncovering the interactive effect. An implication is that the burden, to some extent, falls on an experiment’s critic to identify the moderating factor and demonstrate that it lacks variance in an experiment’s sample.

3. Comparing Student Samples with Other Samples

We argue that a given sample constitutes only one — and arguably not the critical one — of many considerations when it comes to assessing external validity. Furthermore, a student sample only creates a problem when a researcher 1) fails to model a contingent
constructive, more rigorous that reflect its samples may be and direction spent on a particular that are peculiarly its sample.

ied three distinct, moneigen case - ; process produces - we showed that fact derived from a sed estimate of the end, when there is the treatment is covariate (Z) and cognize the continuum (indeed, any 7 misestimate the the sample is non-parallel covariate Z. ven a representa- cify the treatment model the inter- researcher apprehension with an int sample, even if it covariate Z, will if there is virtually (most none) on the cetera, a researcher degree of vari- within a given sample evaluate whether problem for uncover: An implication that is extent, falls on an if the moderating at it lacks variance.

at Samples

ple constitutes only the critical one - when it comes to . Furthermore, a es a problem when model a contingent causal effect (when there is an underlying heterogeneous treatment effect), and 2) the students differ from the target population with regard to the distribution of the moderating variable. This situation, which we acknowledge does occur with nontrivial frequency, leads to the question of just how often student subjects empirically differ from representative samples. The greater such differences, the more likely problematic inferences will occur.

Kam (2005) offers some telling evidence comparing student and nonstudent samples on two variables that can affect information processing: political awareness and need for cognition. She collected data from a student sample using the same items that are used in the American National Election Study's (ANES's) representative sample of adult citizens. She found that the distributions for both variables in the student sample closely resemble those in the 2000 ANES. This near-identical match in distribution allowed Kam to generalize more broadly results from an experiment on party cues that she ran with the student subjects.

Kam (2005) focuses on awareness and need for cognition because these variables plausibly moderate the impact of party cues; as explained, in comparing student and non- student samples, one should focus on possible differences that are relevant to the study in question. Of course, one may nonetheless wonder whether students differ in other ways that could matter (e.g., Sears 1986, 520). This requires a more general comparison, which we undertake by turning to the 2006 Civic and Political Health of the Nation Dataset (collected by CIRCLE) (for a similar exercise, see Kam et al. 2007).

These data consist of telephone and web interviews with 2,232 individuals aged 15 years and older living in the continental United States. We limited the analysis to individuals aged 18 years and older. We selected all ostensibly politically relevant predispositions available in the data and then compared individuals currently enrolled in college against the general population. The Web Appendix contains question wording for each item.

As we can see from Table 4.1, in most cases, the difference in means for students and the nonstudent general population are indistinguishable from zero. Students and the nonstudent general population are, on average, indistinguishable when it comes to partisanship (we find this for partisan direction and intensity), ideology, importance of religion, belief in limited government, views about homosexuality as a way of life, contributions of immigrants to society, social trust, degree of following and discussing politics, and overall media use. Students are distinguishable from nonstudents in religious attendance, level of political information as measured in this dataset, and specific types of media use. Overall, however, we are impressed by just how similar students are to the nonstudent general population on key covariates often of interest to political scientists.

In cases where samples differ on variables that are theorized to influence the size and direction of the treatment effect, the researcher should, as we note previously, model the interaction. The researcher might also consider cases where students – despite differing on relevant variables – might be advantageous. In some situations, students facilitate testing a causal proposition. Students are relatively educated, in need of small amounts of money, and accustomed to following instructions (e.g., from professors) (Guala 2005, 33–34). For these reasons, student samples may enhance the experimental realism of experiments that rely on induced value theory (where monetary payoffs are used to induce preferences) and/or involve relatively complicated, abstract instructions (Friedman and Sunder 1994, 39–40).

19 Available at http://faculty.wcas.northwestern.edu/~jnd260/publications.html.
20 The measure of political information in this dataset is quite different from that typically found in the ANES; it is heavier on institutional items and relies more on recall than recognition.
21 We suspect that this explains why the use of student subjects seems to be much less of an issue in experimental economics (e.g., Guala 2005).
Table 4.1: Comparison of Students versus Nonstudent General Population

<table>
<thead>
<tr>
<th></th>
<th>Students</th>
<th>Nonstudent Population</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Partisanship</td>
<td>0.47</td>
<td>0.45</td>
<td>ns</td>
</tr>
<tr>
<td>(0.02)</td>
<td>(0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ideology</td>
<td>0.50</td>
<td>0.52</td>
<td>ns</td>
</tr>
<tr>
<td>(0.01)</td>
<td>(0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Religious attendance</td>
<td>0.56</td>
<td>0.50</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>(0.02)</td>
<td>(0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Importance of religion</td>
<td>0.63</td>
<td>0.62</td>
<td>ns</td>
</tr>
<tr>
<td>(0.02)</td>
<td>(0.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Limited government</td>
<td>0.35</td>
<td>0.33</td>
<td>ns</td>
</tr>
<tr>
<td>(0.03)</td>
<td>(0.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homosexuality as a way of life</td>
<td>0.60</td>
<td>0.62</td>
<td>ns</td>
</tr>
<tr>
<td>(0.03)</td>
<td>(0.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contribution of immigrants to society</td>
<td>0.62</td>
<td>0.63</td>
<td>ns</td>
</tr>
<tr>
<td>(0.03)</td>
<td>(0.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social trust</td>
<td>0.34</td>
<td>0.33</td>
<td>ns</td>
</tr>
<tr>
<td>(0.03)</td>
<td>(0.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Follow politics</td>
<td>0.68</td>
<td>0.65</td>
<td>ns</td>
</tr>
<tr>
<td>(0.02)</td>
<td>(0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discuss politics</td>
<td>0.75</td>
<td>0.71</td>
<td>ns</td>
</tr>
<tr>
<td>(0.01)</td>
<td>(0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Political information (0-6 correct)</td>
<td>2.53</td>
<td>1.84</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>(0.11)</td>
<td>(0.07)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Newspaper use (0-7 days)</td>
<td>2.73</td>
<td>2.79</td>
<td>ns</td>
</tr>
<tr>
<td>(0.14)</td>
<td>(0.11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>National television news (0-7 days)</td>
<td>3.28</td>
<td>3.63</td>
<td>&lt;.05</td>
</tr>
<tr>
<td>(0.15)</td>
<td>(0.10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>News radio (0-7 days)</td>
<td>2.47</td>
<td>2.68</td>
<td>ns</td>
</tr>
<tr>
<td>(0.16)</td>
<td>(0.11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web news (0-7 days)</td>
<td>3.13</td>
<td>2.18</td>
<td>&lt;.01</td>
</tr>
<tr>
<td>(0.16)</td>
<td>(0.10)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall media use</td>
<td>2.90</td>
<td>2.83</td>
<td>ns</td>
</tr>
<tr>
<td>(0.09)</td>
<td>(0.06)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes: ns, not significant. Weighted analysis. Means with standard errors in parentheses. See the Web Appendix for variable coding and question text.


The goal of many of these experiments is to test theory, and, as mentioned, the match to the theoretical parameters (e.g., the sequence of events if the theory is game theoretic) is of utmost importance (rather than mundane realism).

Alternatively, estimating a single treatment effect on a student sample subject pool
can sometimes make it more difficult to find effects. For example, studies of party cues examine the extent to which subjects will follow the advice given to them by political parties. Strength of party identification might be a weaker cue for student subjects, whose party affiliations are still in the formative stages (Campbell et al. 1960). If this were the case, then the use of a student sample would make it even more difficult to discover party cue effects. To the extent that party cues work among student samples, these likely underestimate the degree of cue taking that might occur among the general population, whose party affiliations are more deeply grounded. Similarly, students seem to exhibit relatively lower levels of self-interest and susceptibility to group norms (Sears 1986, 524), meaning that using students in experiments on these topics increases the challenge of identifying treatment effects.\(^2\)

Finally, it is worth mentioning that if the goal of a set of experiments is to generalize a theory, then testing the theory across a set of carefully chosen convenience samples may even be superior to testing the theory within a single random sample.\(^3\) A theory of the moderating effect of attitude crystallization on the effects of persuasive communications might be better tested on a series of different samples (and possibly different student samples) that vary on the key covariate of interest.

Researchers need to consider what particular student sample characteristics might lead a causal relationship discovered in the sample to differ systematically from what would be found in the general population. Researchers then need to elaborate on the direction of the bias: the variation might facilitate the assessment of causation, and/or it might lead to either an overestimation or an underestimation of what would be found in the general population.

4. Conclusion

As mentioned, political scientists are guilty of a "near obsession" with external validity (McDermott 2002, 334) and this obsession focuses nearly entirely on a single dimension of external validity – who is studied. Our goal in this chapter has been to situate the role of experimental samples within a broader framework of how one might assess the generalizability of an experiment. Our key points are as follows:

- The external validity of a single experimental study must be assessed in light of an entire research agenda and in light of the goal of the study (e.g., testing a theory or searching for facts).
- Assessment of external validity involves multiple dimensions, including the sample, context, time, and conceptual operationalization. There is no reason per se to prioritize the sample as the source of an inferential problem. Indeed, we are more likely to lack variance on context and timing because these are typically constants in the experiment.
- In assessing the external validity of the sample, experimental realism (as opposed to mundane realism) is critical, and there is nothing inherent to the use of student subjects that reduces experimental realism.
- The nature of the sample – and the use of students – matters in certain cases. However, a necessary condition is a heterogeneous (or moderated) treatment effect. Then the impact depends on the following:
  - If the heterogeneous effect is theorized, then the sample only matters if there is virtually no variance on the moderator. If there is even scant variance, then the treatment effect will not only be correctly estimated but may also be estimated with greater confidence. The suitability of a given sample can be

\(^2\) As explained, students also tend to be more susceptible to persuasion (Sears 1986). This makes them a more challenging population on which to experiment if the goal is to identify conditions where persuasive messages fail (e.g., Druckman 2001).

\(^3\) Convenience samples might be chosen to represent groups that are high and low on a particular covariate of interest. This purposive sampling might yield more rewards than using a less informative random sample.
assessed (e.g., empirical variance can be analyzed). • The range of heterogeneous, nontheorized cases may be much smaller than is often believed. Indeed, when it comes to a host of politically relevant variables, student samples do not significantly differ from nonstudent samples. • There are cases where student samples are desirable because they facilitate causal tests or make for more challenging assessments.

Our argument has a number of practical implications. First, we urge researchers to attend more to the potential moderating effects of the other dimensions of generalizability: context, time, and conceptualization. The past decade has seen an enormous increase in survey experiments due in no small way to the availability of more representative samples. Yet, scholars must account for the distinct context of the survey interview (e.g., Converse and Schuman 1974; Zaller 1992, 28). Sniderman, Brody, and Tetlock (1991) elaborate that "the conventional survey interview, though well equipped to assess variations among individuals, is poorly equipped to assess variation across situations" (265). Unlike most controlled lab settings, researchers using survey experiments have limited ability to introduce contextual variations.

Second, we encourage the use of dual samples of students and nonstudents. The discovery of differences should lead to serious consideration of what drives distinctions (i.e., what is the underlying moderating dynamic, and can it be modeled?). The few studies that compare samples (e.g., Gordon et al. 1986; James and Sonner 2001; Peterson 2001; Mintz, Redd, and Vedlitz 2006; Depositario et al. 2009; Henrich, Heine, and Norenzayan 2009), although sometimes reporting differences, rarely explore the nature of the differences. 24 When dual samples are not feasible, researchers can take a second-best approach by utilizing question wordings that match those in general surveys (thereby facilitating comparisons).

Third, we hope for more discussion about the pros and cons of alternative modes of experimentation, which may be more amenable to using nonstudent subjects. Although we recognize the benefits of using survey and/or field experiments, we should not be overly sanguine about their advantages. For example, the control available in laboratory experiments enables researchers to maximize experimental realism (e.g., by using induced value or simply by more closely monitoring the subjects). Similarly, there is less concern in laboratory settings about compliance × treatment interactions that become problematic in field experiments or spillover effects in survey experiments (Transue, Lee, and Aldrich 2009; also see Sinclair’s chapter in this volume). The increased control offered by the laboratory setting often affords greater ability to manipulate context and time, which, we argue, deserves much more attention. Finally, when it comes to the sample, attention should be paid to the nature of any sample and not just student samples. This includes consideration of nonresponse biases in surveys (see Groves and Peytcheva 2008) and the impact of using “professional” survey respondents that are common in many web-based panels. 25 In short, the nature of any particular sample needs to be assessed in

---

24 For example, Mintz, Redd, and Vedlitz (2006) implemented an experiment, with both students and military officers, about counterterrorism decision making. They found that the two samples significantly differed, on average, in the decisions they made, the information they used, the decision strategies they employed, and the reactions they displayed. Mintz et al. conclude that "student samples are often inappropriate, as empirically they can lead to divergence in subject population results" (759). We would argue that this conclusion is premature. Although their results reveal differences, on average, between the samples, the authors leave unanswered why the differences exist. Mintz et al. speculate that the differences may stem from variations in expertise, age, accountability, and gender (756). A thorough understanding of the heterogeneity in the treatment effects (which, as explained, is the goal of any experiment) would, thus, require exploration of these moderators. Our simulation results suggest that even if the student sample exhibited limited variation on these variables, it could have isolated the same key treatment dynamics that would be found in the military sample.

25 The use of professional, repeat respondents raises similar issues to those caused by repeated use of participants from a subject pool (e.g., Stevens and Ash 2001).
Students as Experimental Participants

light of various trade-offs, including consideration of an experiment’s goal, costs of different approaches, and other dimensions of generalizability.

We have made a strong argument for the increased usage and acceptance of student subjects, suggesting that the burden of proof be shifted from the experimenter to the critic (also see Friedman and Sunder 1994, 15). We recognize that many will not be persuaded; however, at the very least, we hope to stimulate increased discussion about why and when student subjects may be problematic.

References


Sears, David O. 1986. “College Sophomores in the Laboratory: Influence of a Narrow Data Base on Social Psychology’s View of Human


